1. Tuning Efforts

We first performed some preprocessing on the complete training data and then divided it into two parts:

1. Training set (80%)
2. Cross-Validation set (20%)

Then we started building training our models on the Training set and then evaluated these models based on their accuracy on the cross-validation dataset. Our criteria of selection between models were that the higher the accuracy, the better the model.

a) HMM Tagger

For the HMM Tagger, we tried to tune the model on the parameter **“estimator”** which is a function that maps a conditions’ frequency distribution to its probability distribution. We tried the following values for this parameter.

MLEProbDist (Lidstone with gamma = 0.1) - default
LidstoneProbDist
LaplaceProbDist
ELEProbDist
WittenBellProbDist

We found that accuracy increased substantially when we replaced the default function of MLEProbDist with any of these other functions.

Estimator	**Cross Validation Accuracy**
LidstoneProbDist	0.00119
MLEProbDist (default)	0.34376
LaplaceProbDist	0.76962
ELEProbDist	0.80930
WittenBellProbDist	0.87446

WittenBellProbDist was giving the maximum accuracy on the cross-validation set and hence we chose this parameter as our final value of ***estimator*** for the HMM tagger.

b) Brill Tagger

For the Brill Tagger, we tried to tune the model on the parameters **“template”** which represents what templates should be used in training, and **“max_rules”** which represents the maximum number of rules instances to create.

We tried the following functions for the template parameter:

nltkdemo18()
nltkdemo18plus()
fntbl37()
brill24()

We tried the following values for the max_rules parameter.

10
50
75
100

Since we have 4 possible values for the first parameter and 4 possible values for the second parameter, we evaluated a total of 4*4= **16 possible models.**

Template	**max_rules**	**Cross Validation Accuracy**
nltkdemo18()	10	0.3890
	50	0.4857
	75	0.5252
	100	0.5395
nltkdemo18plus()	10	0.4111
	50	0.5033
	75	0.5433
	100	0.5633
fntbl37()	10	0.5040
	50	0.6661
	75	0.6962
	100	0.7169
brill24()	10	0.5040
	50	0.6617
	75	0.6986
	100	0.7183

As we can see both fntbl37() and brill2() template rules are giving very high accuracies as compared to other template rules. Also, we can see that the more the number of max_rules, the higher the corresponding accuracy on the cross-validation set.

Since brill24() with 100 rules is giving the highest accuracy, we choose these 2 values as our optimal parameters for the Brill Tagger.

Note: We have made the assumption that accuracy is our only parameter for comparison. If we were also considering the time taken to train, give prediction then max_rules=75 would have been better overall as it has the benefit of taking lesser time, and loss in the accuracy is also not that significant as compared to the tagger that uses 100 rules but is comparatively slower.

2. Tagger Accuracies

After we have built our HMM tagger and Brill Tagger models with the optimal parameters found during hyperparameter tuning, we get the following accuracy values for the 4 combinations of models and testing files.

Tagger	**Test File**	**Accuracy**
HMM	test.txt	**0.8625**
HMM	test_ood.txt	**0.8183**
Brill	test.txt	**0.7079**
Brill	test_ood.txt	**0.5915**

3. Tagger Comparison

We can observe that HMM taggers are giving higher accuracies as compared to Brill Taggers. **The difference in accuracy between these taggers on the in-domain test file is 0.1546 while the difference in the case of out-of-domain test files is even higher at 0.2268.**

This indicates that apart from giving higher accuracies, **the HMM tagger is generalizing much better than the Brill Tagger.** We can say that the Brill tagger has some sort of overfitting because its performance on a test file with sentences it has not seen previously is coming to a lot lesser as compared to sentences that are similar to the ones it had seen while training. This can be partly explained by the fact that HMM is a stochastic approach that incorporates probability calculations to decide the POS tag of a word whereas Brill tagger is a rule-based approach. So if the training data leads to the learning of specific rules, and those rules are not applicable to the out-of-domain test sentences, then the performance of Brill tagger will degrade.

Overall, HMM tagger would be preferred over the Brill tagger for the data we are using in this assignment.

4. Error Analysis

For the HMM Tagger, we noticed some errors such as

1. **Inability to distinguish between Singular Nouns (NN) and Plural Nouns (NNS).** For example, the HMM tagger wrongly tags the word “conditions” as NN instead of NNS in line 3623 (test.txt). Another example is for the word “seasonings”, the tagger wrongly predicted NN instead of NNS. This could be due to the fact that there were more singular noun instances than plural noun instances in the training data.

2. **Similar confusion exists between the tags Proper noun (NNP) and Singular Nouns (NN).** For example, the HMM tagger wrongly predicts the word “Arizona” as NN instead of NNP.

3. **Inability to distinguish between Adverbs and Adjectives.** For example, the HMM tagger wrongly tags the word “longer” as Adjective (JJR) instead of Adverb (RBR) in the sentence “This should take no longer than 30 seconds ”**.** This is understandable as it is difficult for humans as well to decide whether a word is describing an action (adverb) or a noun (adjectives) in similar sentences.

4. **Wrong prediction of some tags as Particle (RP) instead of Preposition (IN).** For example, the word “off” is wrongly predicted as a particle instead of a preposition. We were expecting the tagger to make these errors as words like “off”, “to” can be both a particle and a preposition according to the context in which they are used.

For the Brill Tagger, we noticed that:

1. **Unlike the HMM tagger, the Brill tagger did not make the errors of wrongly predicting** words like “off” as **particles instead of prepositions.** For example, Brill tagger correctly predicted the word “off” as a preposition in the sentence “Take the quinoa off the heat and let stand covered for 5 minutes.”

2. **Unlike the HMM tagger, the Brill tagger was making much more errors of wrongly predicting a word as Singular Noun (NN) instead of Adjective (JJ).** For example, the word “central” was wrongly predicted as Noun (NN) instead of Adjective (JJ) in the phrase “The Salt and Verde rivers of **central** Arizona were.. ”

3. **In general, the Brill tagger was making more errors of wrongly identifying Modals (MD) as Adjectives (JJ) or Noun (NN).** For example, in the sentence “Whenever you are presented with something that you don’t want to do or **would** consider….”, the brill tagger wrongly predicted “would” as NN instead of MD. **These type of errors was not observed in the case of HMM tagger.**

4. **Other types of errors like the confusion between Singular (NN) and Plural (NNS) nouns, between Proper Noun (NNP) and Singular Nouns (NN) etc. are common for both the taggers.**